Submission of Small Variations to dbSNP

Version 3.4, Jun 7, 2014
CURRENTLY UNDER REVISION

This page provide guidelines for dbSNP submission process (PDF version) .

Submission Related Email accounts:

- Submissions to snp-sub@ncbi.nlm.nih.gov
- Updates to snp-update@ncbi.nlm.nih.gov
- Questions, etc. to snp-admin@ncbi.nlm.nih.gov

Variation Size and Data Submission Limitations

- Submit variations >50 nucleotides in length to the Database of Genomic Structural Variation (http://www.ncbi.nlm.nih.gov/dbvar)
- dbSNP does not accept synthetic mutations
- dbSNP does not accept variations ascertained from cross-species alignments and analysis
- dbSNP does not accept personal human data due to current NIH policy unless the participant is enrolled in a study with institutional oversight

Submission for Clinical Variations/Mutations Related to a disease or other phenotypes

- ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/) now accepts submissions of clinically related variations/mutations, and we encourage our submitters to submit their clinically related data there. Your submissions to ClinVar will be processed, assigned ClinVar accessions (SCV), and will be accessioned with novel variant locations in dbSNP or dbVar as appropriate.
- dbSNP and ClinVar will continue to maintain the Human Variation Batch Submission site (http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/VarBatchSub.cgi) as a web-based tool that can be used to submit clinical assertion updates for existing variations.

Submission for Non Clinical Variations ***IMPORTANT UPDATES***

Note: The flat-file format may not be supported for future submissions. Please see details below for alternative VCF format and Excel template.

dbSNP has recently upgraded its submission processing system to accept VCF (Variant Call Format), so we are encouraging all of our users to switch from dbSNP's old flat file submission format to the new VCF submission format. dbSNP's VCF format is very flexible and can be altered to fit the requirements of the data being submitted, so you can use it to submit many kinds of common variations as well as their associated genotypes and annotation. The dbSNP VCF format can be used to submit both large and small scale submissions over multiple populations.

One of the benefits of using the VCF format is that it allows a submitter to submit an "asserted position" rather than flanking sequence as a means to locate variations. An "asserted position" is a statement, or assertion, based on experimental evidence that a variant is located at a particular position. Using asserted positions allows for much greater accuracy in variation mapping than does flanking sequence.

We prefer that all variant asserted positions are submitted on a sequence accession that is part of an assembly housed in the NCBI Assembly Resource.

NOTE 1: *Those variant positions reported on a sequence that is part of an assembly housed in the NCBI Assembly Resource will receive a submitted SNP (ss) number, and a Reference SNP (rs or RefSNP) number. Variations that are assigned a refSNP number are distributed as part of dbSNP, which allows the reported variation to appear on maps or graphic representations of the assembly, and be integrated with NCBI's other resources like Gene, ClinVar, dbGAP or PubMed.*

NOTE 2: *Those variant positions reported on a sequence that does not yet align to an assembly in NCBI's Assembly Resource either because there is not yet an assembly to which the sequence aligns, or because the submitted sequence aligns to a gap in an existing assembly, the variant will be assigned an ss number only. This means that the reported variation will NOT appear on maps or graphic representations of the assembly, and will NOT be integrated with NCBI's other resources. The ss will, however, be reported on the 'Submitted SNP' web report, will be available for search using dbSNP homepage's 'ID search' tool, and will be made available on FTP site for download.*
If, however, at some future date a new assembly is created or an old assembly is updated such that the reported variant sequence aligns to an assembly in the NCBI Assembly Resource, the reported variant will be assigned an rs number at that time, which will allow it to be distributed as part of dbSNP, appear on maps or graphic representations of the assembly, and be integrated with other NCBI resources.

NOTE 3: *Those variant positions reported on a sequence that is only known through an assay that provides just the variant and flanking sequence will be assigned ss numbers only. These ss numbers will be reported in the 'Submitted SNP' web report, will be made available on FTP site for download, and will be available for search using the dbSNP homepage 'ID search' tool. Assay variations will be accessible in dbSNP as archived data until such time as an assembly is available that will allow mapping by BLAST. dbSNP cannot predict when the mapping by BLAST will be completed â€“ it could be delayed by months or possibly years. SS numbers can be used in publications describing these assay variants.*

*If you do not know if your sequence is part of an assembly housed in the dbSNP Assembly Resource

If you do not know if your sequence is part of an assembly housed in the dbSNP Assembly Resource, contact dbSNP at snp-sub@ncbi.nlm.nih.gov*

General information about the VCF format and detailed instructions for creating a VCF formatted submission are available at:
http://www.ncbi.nlm.nih.gov/projects/SNP/docs/dbSNP_VCF_Submission.pdf

A VCF submission template and VCF submission examples are available at:
http://www.ncbi.nlm.nih.gov/projects/SNP/docs/vcf_template.xlsx

Table of Contents

Quick Start

Ready to submit data to dbSNP? Here are some examples of the different sections you can include in the submission file and brief instructions for getting your data into dbSNP.

The basic submission steps:

1. Get a handle assignment from NCBI if your lab doesn't already have one. Send your handle request to snp-admin@ncbi.nlm.nih.gov or use the online handle request form.
2. Prepare a submission file with your data and send it to snp-sub@ncbi.nlm.nih.gov . Several submission scenarios and their respective file components are provided as a guide.
3. You will receive a submission report from NCBI indicating what was loaded into the database, and a list of error or warning messages if problems were encountered while processing your submission file.
4. Resubmissions of corrected files (returned by NCBI because of excessive errors) should be sent to snp-update@ncbi.nlm.nih.gov .

Back to Table of Contents

Database Organization

The purpose and scope of dbSNP

dbSNP is a public database of single nucleotide polymorphisms (SNPs). The data can be from any species, and from any part of a particular genome. SNPs linked to known genes or expressed DNA segments (ESTs) will be particularly useful in the database. Since many of NCBI's resources are gene or map-oriented, SNPs from these regions of the genome will be the first to be integrated to other NCBI resources.

SNPs exist at defined positions within genomes and can be used for gene mapping, defining population structure, and performing functional studies. dbSNP has been designed to include a broad collection of simple genetic polymorphisms such as single-base nucleotide substitutions, small-scale multi-base deletions or insertions, retroposable element insertions and microsatellite repeat variation. Once described, these polymorphisms exist as a public resource for future research, as dbSNP entries record the sequence information around the polymorphism, the specific experimental conditions necessary to perform an experiment, and frequency information by population or individual genotype.

This document describes the procedures for submitting and updating information in dbSNP, and the format for all of the above data for the SNP database maintained by the National Center for Biotechnology Information (NCBI). Note that dbSNP takes the looser 'variation' definition for SNPs, so there is no requirement or assumption about minimum allele frequencies for the polymorphisms in the database.

Back to Table of Contents

Data elements of a submission

Each submission to dbSNP will include some subset of the following items:

- the observed alleles at a particular locus (required).
- the flanking sequence that surrounds the mutation (required).
- genetic map information.
- the experimental method(s) used to assay the variation and their respective protocols and conditions (required).
- population-specific frequency information.
- Individual-specific genotype information
- relevant publications that document the details of the methodologies or populations or both.
- a pointer to a companion dbSTS or GenBank record (required).
- known genes in the region.
- Synonyms for a submitter's SNP ID used in the submission.
- Validation information to describe the quality of the frequency information.

Back to Table of Contents

Resource Integration

Figure 1 illustrates the components of a SNP submission: mutation data, methodologies / experimental conditions, contact information, variation data, and associated entries in other NCBI resources. These components may be related in several ways: as elements within the dbSNP schema; as pointers or feature annotations between separate NCBI resources; or as links between dbSNP and other databases external to NCBI. This collection of information is called a submitted SNP record, and it may be referenced in two ways. First, it may be identified by the name provided by the submitter using the format **HANDLE | ID** (e.g. EXAMPLE | SICKLE01). Alternatively, it may be referred to by the NCBI-assigned submitted-snp accession number which has the format **NCBI | ss<NCBI ASSAY ID>** (e.g. NCBI | ss335). The prefix 'ss' is always lower-case.

Within dbSNP

MUTATION DATA: SNP records contain information on the specific alleles and the flanking sequence that surrounds the mutation.

COLLECTION METHODS: Descriptions of the assay technique used to type the SNP are recorded.

SUBMITTER DATA: Contact information is maintained for each lab director and the individual submitter of each batch of records. Bibliographic data for unpublished or in-press citations are recorded.

VARIATION DATA: the database contains all frequency formation provided by population, and genotype information provided for individuals. Populations are defined by the submitter. Individuals may be sub-classified by population or sample frame.

Between NCBI resources

MUTATION DATA: the PCR protocol, primers and buffer conditions for a SNP are stored in a separate entry in dbSTS. This information may be submitted either prior to, or simultaneously with the SNP submission. Other sequence data can be linked to a SNP record by supplying a GenBank accession number.

COLLECTION METHODS: Published citations are referred to with a PubMed ID.

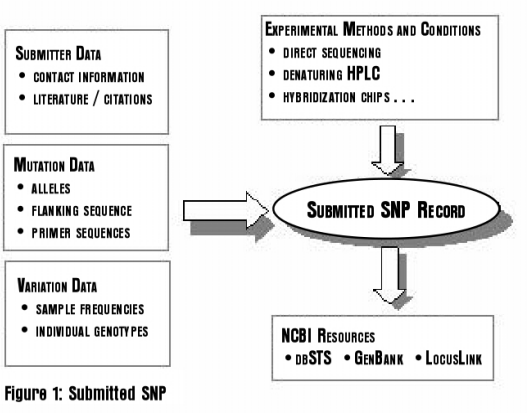

Figure 1: Submitted SNP

Back to Table of Contents

Reference SNP reports and submitter reports have different identifiers in dbSNP

When two submitted SNP records refer to the same location in the genome, records will be related as in Figure 2.It is anticipated that multiple labs may submit information on the same SNPs as new techniques are developed to assay variation, or new populations are typed for frequency information. These Reference SNP records will provide a summary list of submitter records in dbSNP and a list of external resource and database links as illustrated in Figure 2. Reference SNP identifiers will also be exported as standardized features for annotation in other NCBI resources. In this scheme the identifier can be used to retrieve summary information on all the known variation at the locus, a list of the specific reports that characterize a SNP, and links to other NCBI resources.

Reference SNP cluster 'rs' ID's are created by NCBI during periodic 'builds' of the database. Reference SNP clusters define a non-redundant set of markers that are used for annotation of reference genome sequence and integration with other NCBI resources. Novel submissions at new positions in genome sequence will instantiate a new refSNP cluster. New submissions that match existing data will be merged into an existing refSNP cluster. A reference SNP cluster record has the format NCBI | rs[NCBI SNP ID] where 'rs' is always lower case.

To review, SNPs are indexed by two different accession numbers in dbSNP: the **HANDLE | ID** / **NCBI | ssASSAY ID** forms which refer to an individual submission record (Figure 1), and the **NCBI | rsSNP ID** form which refers to the abstracted SNP (Figure 2) and all associated records.

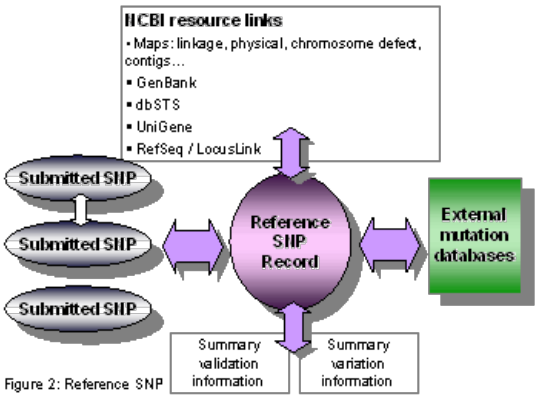

Figure 2: Reference SNP

More information about identifiers in the database may be found in A Note Regarding Identifiers.

Back to Table of Contents

The report of the experimental conditions and molecular context of a SNP is separate from a report of frequency information.

dbSNP distinguishes a report of how to assay a SNP (type SNPASSAY below) from the use of that SNP with individuals and populations (types SNPINDUSE and SNPPOPUSE below). This separation simplifies some issues of data representation. However, these initial reports describing how to assay a SNP will often be accompanied by SNP experiments measuring allele occurrence in individuals and populations. Note that SNP experiments might be performed at a later time, and possibly contributed by labs other than the one who provided the original submission.

There are two meanings for 'population' used in this document. One is the more formal, as would be used by population geneticists. The other is simply to stand for the group of individuals whose DNA was pooled in an experiment. Both are treated the same.

Back to Table of Contents

Database Policies and Administration

A Handle for the submitter is required

Each laboratory will be assigned a "handle" that has multiple uses within the submission format. These handles will be assigned by NCBI during early contacts with each submitter. The handle might be an acronym, or a shorted name of a submitter or large center. This "handle" will allow submissions to be associated with laboratories independent of the details of who is handling a particular set of submissions from that laboratory. Request for a handle online or sent to snp-admin@ncbi.nlm.nih.gov the following information:

```
HANDLE:        A suggested short abbreviation or acronym to identify the lab
NAME:          Name of the lab chief or principal investigator
FAX:           Include area code (and country code if outside of the USA)
```

```
TEL:          Include area code (and country code if outside of the USA)
EMAIL:        Address for lab chief or P.I.
LAB:          Name of lab, or Lab Chief if private lab
INST:
ADDR:         Complete mailing address
```

Submitter information vs. Batch contact information

Contact information for a lab chief will be collected at the time the lab's handle is assigned. This information will be shown on all SNP reports associated with the submitter's handle. Each batch of submissions will also have a contact information block that will be shown in batch summary reports. This information will be used by NCBI if the submitter of a particular batch of data has to be reached to answer questions. Changes in a lab chief's contact information can be made by notifying NCBI at snp-admin@ncbi.nlm.nih.gov. This way, the lab chief stays associated with his/her data in the case of a move to a new institution.

After a user has a handle, SNP entries may be submitted by email to "snp-sub@ncbi.nlm.nih.gov". A special tagged flat file input format (see below) has been designed for this data, to allow it to be submitted as one or more text files in this manner.

Changes to previous submissions should be sent to "snp-update@ncbi.nlm.nih.gov". The same file input format is used for updates as for submissions.

Back to Table of Contents

Getting help

If you have questions about the file format for submissions to dbSNP, or about the submission process itself, please contact "info@ncbi.nlm.nih.gov" and a member of the support staff will get back to you or pass your question on to snp-admin for a response.

Back to Table of Contents

Submitting sequence information with a SNP report

If the sequences on which an STS based SNP assay is based have not been submitted to GenBank, they must either be submitted to dbSTS, before the SNP assays are submitted, or they may be submitted simultaneously, with a minimum of duplicated data. If assays are based on data not in GenBank that cannot be submitted to STS, please contact NCBI for instructions. If this simultaneous submission is done, the STS accession will be automatically added to the SNP assay data. The SNP assay is linked to the STS sequence by using the STS line in the SNP assay report.

Back to Table of Contents

"Hold until published" (HUP) policies

Submitters should note that dbSTS and dbSNP differ in their respective hold until published, or "HUP" policies. Submissions to dbSTS, including the simultaneous submissions discussed above, can be withheld from public view until the accession number is published. dbSNP records, however, will be available for public inspection when the submission process is complete, even in the case of simultaneous dbSNP/dbSTS submissions. STS submissions that require HUP treatment should be submitted separately, and prior to the SNP submission.

Once your data are ready to submit or update, email them to "snp-sub@ncbi.nlm.nih.gov" or "snp-update@ncbi.nlm.nih.gov" as described above.

Back to Table of Contents

Important note regarding Identifiers

There are these "flavors" of identifiers to keep in mind:

- **Public identifiers in other NCBI databases**. These identifiers are keys from other databases that permit the retrieval of requested records. Two examples relevant to dbSNP are
  - PMIDs, for journal articles, and
  - GenBank ACCESSIONs pointing within dbSTS.
- The abstracted **NCBI Reference SNP Cluster Identifier**, which has the form:

  ```
  NCBI|rs<number>  EXAMPLE:   NCBI|rs12345
  ```

  and will be assigned to the abstraction tracked by dbSNP which is the position in the idealized genome where the variation can be assayed. See Abstract and Submitted SNP records have different identifiers in dbSNP for more information.

- The **NCBI submitted SNP Identifier**, which has the form:

  NCBI|ss<number> EXAMPLE: NCBI|ss12345

will be assiged to each submitted SNP report. It is equivalent in function to using the local handle-specific identifier discussed below, but unlike the latter, 'ss' numbers will have a consistent format. See Abstract and Submitted SNP records have different identifiers in dbSNP for more information.

- **Identifiers that point back to other public databases**, especially those of the submitter. dbSNP expects that submitters may have their own, often web based, resources for the public to use. dbSNP will allow linking of submitted assays using multiple such synonyms. These are only expected to be unique when combined with the submitter handle. There are examples within the SNP assay section.

- **Limited or local identifiers**. These are identifiers that might only be unique when combined with the handle **AND** only apply to the dbSNP submission. There are these examples possible within a submission to dbSNP:
  - The identifier used to name a submitted SNP assay. This need not have any meaning outside of dbSNP. Initially, it will be expected to map one-to-one to a single NCBI SNP Identifier, but eventually multiple submitter SNP assays will occasionally map to the same NCBI SNP Identifier. See Abstract and Submitted SNP records have different identifiers in dbSNP for more information.
  - The identifier used to refer to a **method** for assaying SNPs that may be supplied by a submitter.
  - The identifier used to refer to a **population** of individuals that was either used to define a SNP assay or upon which a SNP assay was applied. Note that some population strings will be predefined, or "globally" defined. These may be used by more than one submitter. To remove ambiguity, populations will always be used as <handle>|<population id> with the handle for the globally defined populations being 'NCBI'.
  - The identifier used for an **individual** within a particular defined population. This identifier need only be unique within the combination of <handle>|<population id>, so for example, the integers may be reused for each submitter's population.
  - The identifier used to refer to a batch id. This is simply a name for a set of submitter SNP assays or experiments. Having this name allows for clear reference to the submitted set in communication between NCBI and submitters.

If a submitter is referring to their own **method** or **population**, in the submissions, it is not necessary to add the HANDLE, as it will be assumed that they are referring to their own information. However, it is also possible to refer to another submitters information, in that case adding their "<handle>|" (with the vertical bar separator) would be required.

Back to Table of Contents

Locus Validation and Quality Assurance in dbSNP

Data validation can be maintained for both submitted assay reports and abstracted SNP objects. At the level of an individual submitted assay report, dbSNP provides several fields to assess the quality of the data.

Back to Table of Contents

Submissions to dbSNP

(Note: The flat-file format may not be supported for future submissions. Please see guidelines for the alternative VCF format and Excel template).

Alternative SNP Submissions Scenarios

A number of different "flavors" of submissions are possible to dbSNP. This section presents some of them, so submitters will have a better feel for which of the following detailed sections apply to their situation. Note that some populations and individuals are known to dbSNP, including the NIH NHGRI set called the *Polymorphism Discovery Resource*, which has the Population ID, **`SNP|NIHPDR`**.

SNPs already in dbSTS; SNP assay and use on individuals from a defined population.

1. The submission should contain:
2. The contact information. This block should always be present.
3. Publication section(s).
4. A Method section, unless using a method close enough to one submitted by others so a METHOD_EX (method exception) will suffice to describe the differences.
5. SNP assay section.
6. SNP Use on individuals sections.

Back to Table of Contents

Simultaneous dbSTS/dbSNP assay submission with use on pooled individuals from populations to be defined by the submitter.

Note that on the STS submission page, 'Files' is often used where "Sections" is used herein. This is for historical reasons for dbSTS. Since multiple 'sections' may appear in the same file, the former name is in this document.

1. The submission should contain:
2. The dbSTS submission information, including the handle in the contact block: (see the URL)
   a. `Publication`
   b. `Source`
   c. `Contact`
   d. `Protocol`
   e. `Buffer`
   f. `STS`
   g. `MAP information`

3. A Method section, unless using a method close enough to one submitted by others so a METHOD_EX (method exception) will suffice to describe the differences.
4. SNP assay section.
5. Population description section.
6. SNP use on population section.

Back to Table of Contents

Use of SNP assays defined by others on individuals from populations to be defined by the submitter.

1. The submission should contain:
2. The contact information. This block should always be present.
3. Publication section(s).
4. A Method section, unless using a method close enough to one submitted by others so a METHOD_EX (method exception) will suffice to describe the differences.
5. Population description section.
6. SNP use on individuals section.

Back to Table of Contents

Updating submissions in dbSNP

Updates for the most common situations are described below. Questions about other circumstances can be directed to snp-admin@ncbi.nlm.nih.gov for instructions.

- **Contact information:** changes can be made to an investigator's handle contact information by sending the new information to snp-admin@ncbi.nlm.nih.gov.
- **Publication information:** send an updated publication section to snp-update@ncbi.nlm.nih.gov. The title field must be identical between the original submission and the update information to match the records. If the update is the publication of an "in press" citation, then the PubMedID or MUID may be provided in lieu of the compete reference if it is known. The necessary information will be automatically extracted from the full citation in PubMed.
- **Changing the basic record:** submitters should send the updated sections to snp-update@ncbi.nlm.nih.gov using the same file format as specified for original submissions.

Back to Table of Contents

Reporting sequence variation and mutant alleles

The sequence data captured by the database consists of three elements:

- The sequence 5' to the site of mutation
- The mutation itself (The OBSERVED: line of a SNP assay.)
- The sequence 3' to the site of mutation

This section details the conventions for presenting the mutations observed. Because the methodology for SNP discovery is diverse, a variety of data is expected. For the 5' and 3' sides, it is understood that together they will sum to at least 100 bases. Each taken alone will be 25 bases, minimum size. The standard IUPAC ambiguity characters are permitted in flanking sequence to identify regions of known variation. Used in this context, ambiguous sites would also be SNPs in their own right, and each should have its own separate, simultaneous submission. Ambiguity characters are not to be used to accommodate poor sequencing results. It is understood that for regions of intense variation, the particular haplotype presented in the 5' and 3' regions might be rare.

For SNP assays, each of the alternatives may be separated by a slash ('/'), to denote the alternative alleles observed. The order of each allele in the OBSERVED field does not matter.

Although we normally expect a single slash with two nucleotides, other cases, especially on populations, with many additional slashes can be imagined. Other classes of highly polymorphic markers, such as microsatellites, are expected to have many allelic states. In general, the text in between slashes must be less than 50 characters, and the total text used (on any one OBSERVED: line) must be less than 255 characters.

In general, parenthesis are used to indicate a string which is not actually a nucleotide sequence. The legal formats are:

- Sequence, generally a single base for Single Nucleotide Polymorphisms. So to report both an A and a G at the SNP site:

```
A/G
```

- Some techniques only allow detection of heterozygosity, this would be shown with:

```
(heterozygous)
```

- An indel (insertion/deletion, also called "DIP" for deletion insertion polymorphism) can be shown with a dash ('-'). The position of the dash before or after the "/" does not matter. So to show that at a site which often has a A, that A might be deleted, use:

```
-/A or A/-
```

- An indel of a few bases might occur. These are limited to 50 bases and are generally expected to be less than 20 bases long. So to show a position where an insertion/deletion of GATC might be present or absent, use:

```
-/GATC
```

- To show indel of a repeat element, use a name for the repeat element. For example to show a place where there might be an Alu inserted, use:

```
-/(Alu)
```

- To show microsatellite repeat alleles, use the parenthesized repeat motif followed by observed alleles scored as repeat number. It is understood that the repeating motif may not be exactly conserved in all individuals. The example below illustrates a dinucleotide repeat with 6 alleles:

```
(AT)8/9/10/11/12/13
```

- **ONLY for reporting results on individuals** (SNP:line) **and populations** (SNPCOUNT or SNPFREQ lines), but not for SNP assays, the following may also be used:
  - (homozygous)
  - (indeterminate)
  - (not attempted)
  - (Region deleted)

Back to Table of Contents

The Submission File Format

It is expected that SNP assays will be submitted to dbSNP as batches of dozens to thousands to even hundreds of thousands of entries, with a great deal of redundancy in the citation, submitter and other information. To improve the efficiency of the submission process for this type of data, we have designed a streamlined submission process and data format. These formats are largely based on those used for submission to dbSTS and dbEST.

The following is a specification for flat file formats for delivering SNP assays and the results of use of those SNP assays and related data to the NCBI SNP database. The format consists of colon delineated capitalized tags, followed by data. The data for most fields should appear on the same line as the tag, with no line wrapping. Exceptions to this are clear from the formats as presented, below. In these cases, the data begins on the line following the field tag and can have additional lines. The METHOD_EX is an exception, short text may be on the same line, but may also continue on subsequent lines. For the method and population descriptions, user provided line breaks will be preserved, so additional user defined tagging and formatting can be preserved. Each record (including the last record in the section) should end with a double-bar tag (||) to indicate the end of the record.

NOTE --

Each SNP assay and use in individuals or use in populations submission may reference the Contact data, Publication, Method, and Population submission information. Therefore the submission information for these latter sections must be in the database when the SNP section is entered. This is most easily done by placing these sections at the beginning of a submission file. Once this information has been submitted and entered, it does do not need to be re-submitted for additional SNP assays or use submission sections that have the same Contact, Publication, Method, or Population information.

Back to Table of Contents

Section Types for Submissions to dbSNP

Contact Sections

The following is an example of the valid tags and some illustrative data: (TYPE, HANDLE, and NAME are required.)

```
TYPE: CONT        Entry type - must be "CONT" for contact entries
HANDLE:<handle>   Short name, or handle as supplied by NCBI
NAME:             Name of person who submitted the SNP file.
FAX:              Fax number as string of digits.
TEL:              Telephone number as string of digits.
EMAIL:            E-mail address
LAB:              Laboratory providing SNP.
INST:             Institution name
ADDR:             Address string, comma delineation.
||
e.g.
TYPE:  CONT
HANDLE:EGREEN
NAME:  Eric Green
EMAIL: egreen@wugenmail.wustl.edu
LAB:   Center for Genetics in Medicine
INST:  Washington University School of Medicine
ADDR:  Box 8232, 4566 Scott Avenue, St. Louis, MO 63110, USA
||
```

The TYPE field is obligatory at the beginning of each entry, even if there are multiple entries of a given type in a file. We require the handle, and if this is part of a joint dbSTS submission, the name of a contact person. We would like as many of the fields filled in as possible, to provide complete information to the user for contacting a source for the SNP or further information about it. The handle field in the SNP entries must contain an identical string to the string used for the handle in the contact entry, for automatic matching.

Back to Table of Contents

Publication Sections

The following is an example of the valid tags and some illustrative data: (TYPE, TITLE, YEAR, and STATUS, are required.)

```
TYPE: PUB       Entry type - must be "PUB" for publication entries.
```

```
HANDLE:<handle> Short name, or handle as supplied by NCBI
MEDUID:         Medline unique identifier. Not obligatory,
                include if you know it.
                include if you know it.
PMID:           PubMed unique identifier. Not obligatory,
TITLE:          Title of article.
                (Begin on line below tag, use multiple lines if necessary)
AUTHORS:        Author name, format:  Name,I.I.; Name2,I.I.; Name3,I.I.
                (Begin on line below tag, use multiple lines if necessary)
JOURNAL:        Journal name
VOLUME:         Volume number
SUPPL:          Supplement number
ISSUE:          Issue number
I_SUPPL:        Issue supplement number
PAGES:          Page, format:   123-9
YEAR:           Year of publication.
STATUS:         Status field.
                1=unpublished, 2=submitted, 3=in press, 4=published
||
e.g.
TYPE: PUB
HANDLE: EGREEN
MEDUID:
TITLE:
Human chromosome 7 STS
AUTHORS:
Green,E.
YEAR: 1996
STATUS: 1
||
TYPE: PUB
HANDLE: EGREEN
MEDUID: 96172835
TITLE:
CpG islands of chicken are concentrated on microchromosomes
AUTHORS:
McQueen,H.A.; Fantes,J.; Cross,S.H.; Clark,V.H.;
Archibald,A.L.; Bird,A.P.
JOURNAL: Nat. Genet.
VOLUME: 12
PAGES: 321-4
YEAR: 1996
STATUS: 4
||
```

The TYPE field is obligatory at the beginning of each entry, even if there are multiple entries of a given type in a file. The MEDUID field is a Medline record unique identifier. We do not normally expect you to supply this - we try to retrieve this from our relational version of Medline. The STATUS field is 1=unpublished, 2=submitted, 3=in press, 4=published. The TITLE field is a free format string. The only requirement is that you put an identical string in the CITATION field of the SNP assay or use section, since we will be matching that field automatically against the publications in the publication table and replacing the string with the publication id in the dbSNP table. In practice the handle and title, in combination, must be unique, so submitters may choose any title they wish, even for unpublished citations, as long as it is distinct from other titles that they have used.

Back to Table of Contents

Method Sections

The following is an example of the valid tags and some illustrative data: (all fields required) .

TYPE: METHOD	Entry type - must be "Method" for method entries.
HANDLE: <handle>	Short name, or handle as supplied by NCBI
ID: <local method Identifier>	
METHOD_CLASS: Valid classes are <Sequence, DHPLC, Hybridization, Computation, SSCP, Other, Unknown>	General class of method.
SEQ_BOTH_STRANDS:<YES, NO, NA, UNKNOWN>	Sequenced both strands?
TEMPLATE_TYPE:<DIPLOID, CLONE, OTHER, UNKNOWN>	Was the template DNA used in the assay derived from a clone or from a diploid genomic DNA extraction?
MULT_PCR_AMPLIFICATION: <YES, NO, NA, UNKNOWN>	Independent PCR amplifications tested?
MULT_CLONES_TESTED: <YES, NO, NA, UNKNOWN>	Independent clones tested?
METHOD:	This is multiple lines of free text, however, the line breaks will be preserved and if the submitters use the format
PARAMETER:	Reaction parameters
||	

```
e.g.
TYPE:METHOD
HANDLE:WHOEVER
ID:PROTOCOL-A
METHOD_CLASS: Sequence
SEQ_BOTH_STRANDS: YES
TEMPLATE_TYPE: DIPLOID
```

```
MULT_PCR_AMPLIFICATION: YES
MULT_CLONES_TESTED: NO
METHOD:
PCR reactions were performed with genomic DNA and products were analysed by DNA sequencing.
PARAMETER:
Template: 50 ng genomic DNA
Primer:     each 0.5 uM
dNTPs:     each 0.2 mM
PCR Buffer: 5 ul (10X), Mg 2+ 1.5 mM, Taq Polymerase: 1.25units/ul
||
```

The TYPE field is obligatory at the beginning of each entry, even if there are multiple entries of a given type in a file.

Back to Table of Contents

Population Description Sections

The following is an example of the valid tags and some illustrative data: (All fields required)

```
TYPE:   POPULATION     Entry type - must be "POPULATION" for
                       Population entries.
HANDLE: <handle>       Short name, or handle as supplied by NCBI
ID:     <local Population Identifier>
POP_CLASS: <population geographic class>
MANDATORY:
This free text is a mandatory comment to be displayed
each time any sequence from this population is provided.
This is to be avoided whenever possible, but is added
when consent forms require.
POPULATION:
This is multiple lines of free text, however, the
line breaks will be preserved and if the submitters
use the format
PARAMETER:VALUE
in this text, as much as possible, it will allow
future queries and control.
||
```

Population class:	
central asia	Samples from Russia and satellite Republics, Nations bordering Indian Ocean between East Asia and Persian Gulf regions.
central/south africa	Nations south of Equator, Madagascar & neighboring Island Nations.
central/south america	Samples from Mainland Central and South America, Island Nations of western Atlantic, Gulf of Mexico and Eastern Pacific.
east asia	Samples from Eastern and South Eastern Mainland Asia, Northern Pacific Island Nations.
europe	Samples from Europe north and west of Caucasus Mountains, Scandinavia, Atlantic Islands.
multi-national	samples that were designed to maximize measures of heterogeneity or sample human diversity in a global fashion. Examples OEFNER|GLOBAL and CEPH repository.
north america	All samples north of Tropic of Cancer. This would include defined samples of U.S. Caucasians, African Americans and Hispanics and NCBI|NIHPDR.
north/east africa & middle east	samples collected from North Africa (including Sahara desert), East Africa (south to Equator), Levant, Persian Gulf.
pacific	Samples from Australia, New Zealand, Central and Southern Pacific Islands, Southeast Asian Peninsular/Island Nations.
unknown	Samples with unknown geographic provenience that are not global in nature.
west africa	Sub-Saharan Nations bordering Atlantic north of Congo River, and Central/Southern Atlantic Island Nations.

```
e.g.
TYPE:POPULATION
HANDLE:WHOEVER
ID:YOUR_POP
POP_CLASS: EUROPE
POPULATION:
Continent:Europe
Nation:Some Nation
Phenotype:You name it
||
```

The TYPE field is obligatory at the beginning of each entry, even if there are multiple entries of a given type in a file. The specific fields used above, "Continent", "Nation", and "Phenotype" are for illustrative purposes only. The submitter should choose tags which they judge to be meaningful for their particular population. They also need not use tags, if in their judgement, this would not make sense for their population.

Back to Table of Contents

Individual Description Sections

The following is an example of the valid tags and some illustrative data: (All fields required)

```
TYPE: INDIVIDUAL
IND:handle|loc_pop_id|loc_ind_id|tax_id|sex|breed_structure|ind_grp
SOURCE: src_type|source_name|src_ind_id|loc_ind_grp
PEDIGREE: curator|curator_ped_id|curator_ind_id|ma_ind_id|pa_ind_id
||
```

Individual Submission Data Dictionary

Data element	Data type	Brief Description	Examples	Notes
handle	varchar(64)	dbSNP assigned submission handle	TSC-CSHL	[1]
loc_pop_id	varchar(64)	reference to a submitters population identifier	HapMap-CEU	
loc_ind_id	varchar(64)	submitters individual/sample name	CEPH1331-01	[2]
tax_id	int	terminal tax_id of individual from NCBI taxonomy tree	9606 => Human	[3]
sex	char(1)	sex/gender (optional), male, female, hermaphrodite	M/F/H	
breed_structure	char(1)	I=Inbred, O=outbred, S=structured	I/O/S	
ind_grp	varchar(64)	Broad Grouping of Individual's heritage	European	[4]
src_type	varchar(10)	sample source authority classification	repository or curator or submitter	[5]
source_name	varchar(22)	sample source authority name	Coriell, The Jackson Laboratories, MARC	[6]
src_ind_id	varchar(64)	source authority individual/sample number for loc_ind_id	1331-01, NA1234, Blackie, A/J	[7]
loc_ind_grp	varchar(32)	source's individual group assignment	Western & Northern European, Angus, Affected	[8]
curator	varchar(12)	pedigree curator's name	CEPH,WI, MARC, NCBI	[9]
curator_ped_id	varchar(12)	pedigree identifier in pedigree authority name space	1331	
curator_ind_id	varchar(12)	individual identifier in pedigree authority name space	01	
ma_ind_id	varchar(64)	maternal individual id in pedigree authority name space	05	[10]
pa_ind_id	varchar(64)	paternal individual id in pedigree authority name space	06	[11]

[1]
dbSNP assigned submission handle. Handle request be made online.
[2]
Submitter's sample or individual ID, may be same as src_ind_id
[3]
Use NCBI taxonomy ID
[4]
Optional field to group individuals with similar heritage
[5]
users normally will enter submitter, when known user should also enter repository information see submission example
[6]
When blank src_type becomes submitter and submitting labs Handle becomes source_name
[7]
May be multiple samples from one individual, each sample should be submitted for src_type repository
[8]
details of the Individual/sample grouping should be listed in the population description.
[9]
Unrelated individuals do not need ped infor section however ped id's will be assigned by dbSNP on submission
[10]
Enter 0 for founders
[11]
Enter 0 for founders
[12]
One row for each submitted sample
[13]
May have multiple rows for each sample
[14]
May have multiple rows for each sample

```
Human samples

TYPE: INDIVIDUAL
IND:PERLEGEN|1371|NA06990|9606|F|O|European
SOURCE:repository|Coriell|NA06990|HAPMAP01
SOURCE:curator|CEPH|CEPH1331.02|Utah
PEDIGREE: Coriell|NA06990|NA07050|NA07016
PEDIGREE: CEPH|01|CEPH1331|CEPH1331.02|CEPH1331.15|CEPH1331.14
||
TYPE: INDIVIDUAL
IND:PERLEGEN|1371|NA07050|9606|F|O|European
SOURCE:repository|Coriell|NA07050|Caucasian
SOURCE:curator|CEPH|CEPH1331.15|Utah
PEDIGREE: Coriell|NA07050|0|0
PEDIGREE: CEPH|01|CEPH1331|CEPH1331.15|0|0
||
TYPE: INDIVIDUAL
IND:PERLEGEN|1371|NA07016|9606|M|O|European
SOURCE:repository|Coriell|NA07016|Caucasian
SOURCE:curator|CEPH|CEPH1331.14|Utah
PEDIGREE: Coriell|01|NA07016|0|0
PEDIGREE: CEPH|CEPH1331|CEPH1331.14|0|0
||
```

No Variation Section

This section is used to report no variation in a specified sequence using a particular method and set of samples.

```
TYPE:       NOVARIATION
HANDLE:     <handle>
BATCH:      <local batch id>
```

```
MOLTYPE:     Genomic|cDNA|Mito|Chloro        Molecule type
METHOD:      <local method identifier>
METHOD_EX:   Free text                       variation from given method description
SAMPLESIZE:  <number>                        number of distinct chromosomes examined used as
                                             default value for all records in the batch.
ORGANISM:    scientific name                 as on NCBI taxonomy
STRAIN:      strain name                     (optional)
CULTIVAR:    cultivar name                   (optional)
POPULATION:  <local population identifier>
CITATION:    Title of publication
LINKOUT_URL: Free text (255 char max)        URL to submitter webpage to link local data.
COMMENT:     Free text                       for public. Will be shown with the No Variation report
PRIVATE:     Free text                       note to NCBI for aid in processing.
||
  -  -  -  -  REPEATING FOR EACH STS or SEQUENCE TO BE REPORTED IN THE BATCH   -   -   -   -   -
STS:         <accession> or local-STS-ID     ID for the STS (if applicable). Use <accession> for
                                             records already in dbSTS and <local-STS-ID> for new
                                             STS records with the accompanying STS sections for a
                                             simultaneous STS submission.
ACCESSION:   <accession>[,<ACCESSION>,...]   One or more accession numbers from GenBank. At least
                                             one is required if no STS data/accession is provided.
SAMPLESIZE:   <number>                       Distinct number of chromosomes examined for this sequence.
                                             Default value in batch header will be used if absent here.
COMMENT:     Free text
ASSAY_SEQ:   sequence assayed for variation  The actual sequence examined for possible variation.
||
e.g.
TYPE: NOVARIATION
HANDLE:     OEFNER
BATCH: 99-07-26
MOLTYPE: Genomic
METHOD: DHPLC
SAMPLESIZE: 240
ORGANISM: Homo sapiens
POPULATION:  Global
||
ACCESSION: G42836
ASSAY_SEQ: GTACTGTCTTTACTGGATTATTTCCATTCTCCTTTCCAGAACTCCCCCTGGACAGGGGGA
                GACAGATGTCTGCACTTCTGGACCTCACCAGGCCTCGAACTTTGCTTTTACCCTTTCCAC
                ATAATTATCCTGTCCTGCCACATTCTGAGAGAATTTTCTGGAACGCAGTTCCATGAAGAC
                AGCAAATTTTGCTCAGGACAGAGTCTGGCACACAGTGGGTGCTCAAGCAGCAGCTGCTGA
                ATGGATTCCTCAGCCCTATCTCCCAGCTCTTCAGCCGAGCTGATTCTGCTGTTTGTCCCG
                TTTCTTATGTTATTAATTTCAACCATTATATTTTTTATTTTTGAGAGTTTTGATGATAGA
                GGGAGTTAGAGCTAGTCAAGAGTAGGCCTGAAATATTTAGAAAATGCCTTTGGTCTGGGT
                CCTCAAAGCATTGTGGTTACTTCAGGGATGACACAGGACATGATTTGAGACATTCATATG
||
ACCESSION: G42836
SAMPLESIZE: 18
ASSAY_SEQ: CNCCGCTCCGTGAGTATCCTTNCNCCATCTCCACCCGTGTGCAAGTGTATCCTAGGGGTG
                AAAACCTAGAAGTAGGGTTGCTGTCCGATGCGGCTGAACTGCCCTGCACAGAGGCTGTNC
                CCACGTAGGCGCCTCCAGTGGTGCCCTCACGGAATGGTCAGGCCACTCTTTGCCAAGCCT
||
```

Back to Table of Contents

SNP Assay Section

At the beginning of the section describing the SNP assays there is a header that supplies information that applies to the rest of the section. The required fields in this header are HANDLE, BATCH, MOLTYPE, SAMPLE SIZE and METHOD. If ORGANISM is left out, *Homo sapiens* will be assumed.

```
TYPE:       SNPASSAY                    Entry type, must be "SNPASSAY" for these.
HANDLE:     <handle>                    The submitter and NCBI will agree to a
                                        unique "handle".

BATCH:      <local batch id>            The submitter will name each batch for
                                        ease of communication.  Within a handle, local batch id should be unique.
                                        This is necessary to track each submission for a submitter.

MOLTYPE:    Genomic|cDNA|Mito|Chloro     Since this is so important and could
                                        vary by method, it goes with the header.
                                        If you would like to submit a mixture of
                                        molecular types, please split your submission,
                                        so each contains SNPs assayed using a single
                                        moltype.
METHOD:     <local method Identifier>

METHOD_EX:  Free text                   variation from given method

SUCCESS_RATE:   100%                    Probability that SNP is real, based on validation. Defined as
                                                    1 - false positive rate.
SAMPLESIZE:  <number>                   The number of distinct chromosomes examined in the
                                        course of discovery of the variation.

SYN NAMES:  name[,name,...]             Defines, with a submitter defined label, the
                                        meaning of the synonyms presented on the
                                        "SYNONYM" lines that is allowed with each
                                        SNP assay in the batch.  This ordering and
                                        labeling only applies to this batch.
```

```
                                    For example:
                                    SYN NAMES    SNPid,DnaId,MapDna
ORGANISM:  SCIENTIFIC NAME          as on NCBI taxonomy
STRAIN:    strain or breed name     provide if the sampled germplasm has distinctive properties (e.g. inbred mice,
                                    commercial livestock breeds, or pooled DNA sample for SNP discovery).
                                    Individuals with genotype data referencing variations in this batch may
                                    have different strain or breed attributes, . Those data are provided
                                    separately in the population and pedigree section).


CULTIVAR:  cultivar name            provide if organism is a laboratory cultivar
POPULATION:  <local population identifier>
CITATION:  Title of publication     To match the title of an entry in a
                                    publication section of this submitter.
                                    This field may repeat.
                                    If omitted and a single citation is included
                                    in the batch, the parser will associate the
                                    citation with the assay.
LINKOUT_URL:  Free text (255 char max) URL to the submitter's local website. NCBI requests that
                                    links to data for individual SNP records be formed by the
                                    concatenation of this URL string with the local SNP id.
COMMENT:   Free text                for public, will be shown with each SNP
                                    assay in this batch.
PRIVATE:   Free text                for NCBI to aid in processing
||
     - - - - - Repeating for each SNP Assay   - - - - -
```

(Note: The flat-file format may not be supported for future submissions. Please see guidelines for the alternative VCF format and Excel template).

The SNP_LINK, SYNONYM, SEGREGATES, INDHMZYDET, PCRCONFIRMED, EXPRESSED_SEQUENCE, SOMATIC, COMMENT, METH_FAILURE, and GENENAME fields are optional. One or more of STS or ACCESSION must be supplied. 5'_FLANK and 3'_FLANK are optional if sufficient sequence is specified in 5'_ASSAY and 3'_ASSAY.

```
                                    Description
SNP:  <ID>                          The handle in the header will be associated
                                    with the <ID> provided here, and the combination
                                    must be unique for a particular submitter.
SNP_LINK: <handle>|<ID>,            This field indicates identity between
          [NCBI|ss<ASSAY ID>],      the current submission and a previously
          [NCBI|rs<SNP ID]          reported SNP. This assertion of identity
                                    suspends the usual requirement of 100 b.p. minimum
                                    sequence in the flanking-sequence and assay-sequence
                                    fields discussed below.
SYNONYM:    <ID>[,ID,...]           Other IDs used by the submitter to refer to the SNP.
STS:        <accession> OR Local-STS-ID  Use the Local-STS-ID form only for
                                    simultaneous STS submissions.
                                    This will allow linking by NCBI
                                    with the accession to be assigned
                                    by NCBI.
ACCESSION:  <accession>[,accession,...]  really not as good as an STS,
                                    used if STS is absent.
SAMPLESIZE: <number>                Number of distinct chromosomes examined in the course of
                                    discovery of the SNP. This value will override the value
                                    given in the batch header if present.
SEGREGATES:           YES|NO|UNKNOWN  Has this SNP been shown to "mendelize"?
INDHMZYDET:           YES|NO|UNKNOWN  Were homozygote individuals observed
                                    in the sample?
PCRCONFIRMED:         YES|NO|UNKNOWN  Was polymorphism found on repeat PCR
                                    sample (not an artifact)?
EXPRESSED_SEQUENCE:   YES|NO|UNKNOWN  Is this SNP part of an exon or UTR?
SOMATIC:              YES|NO|UNKNOWN  Is this SNP known to be a somatic mutation?
COMMENT:              Free text

METH_FAILURE:         Free text     This field can be used to add a comment
                                    about problems with an assay, such as
                                    problematic primers. Can be used with
                                    KNOWN_SNP_LINK to report problems with
                                    other assays.
GENENAME:   <gene name>             This to to allow the submitter to
                                    specify a gene name should it be
                                    known.  Obviously, the best name
                                    would be from a controlled set.
                                    (Such as the HUGO set, which can
                                    be _browsed on the web.)
                                    This is a free text field.
LOCUSID:    <number>                Number for the gene assigned in the NCBI
                                    LocusLink database.
LENGTH:     [?|Sequence length]     So software can confirm integrity.
                                    By convention, add 1 (one) for
                                    SNP allele.
                                    For situations where the submissions
                                    are generated by hand, a '?' may be
                                    used and dbSNP will calculate the
                                    length.
5'_FLANK:   <sequence>              Flanking sequence 5' of the assayed
                                    region. Field is required if the
                                    5'_ASSAY is less than 25 b.p. or if
                                    the 5'_ASSAY and 3'_ASSAY fields combined
                                    are less than 100 b.p. Minimum b.p. requirement
                                    is suspended if a valid SNP_LINK field
                                    is provided. White space allowed, and
```

```
                                        will be ignored.
5'_ASSAY:   <sequence>                  Sequence 5' of OBSERVED and detected
                                        by the experiment. White space allowed,
                                        and will be ignored. If less than 25
                                        bases, then 5'_FLANK is also required.
                                        Field may be up to 255 b.p. in size.
                                        If greater than 255 b.p., excess
                                        characters should be put in 5'_FLANK.
OBSERVED:   See the section on reporting SNP variation, above.
ANCESTRAL:  <allele>                    Allele must be from string in OBSERVED field.
3'_ASSAY:   <sequence>                  Sequence 3' of OBSERVED and detected
                                        by the experiment. White space allowed,
                                        and will be ignored. If less than 25
                                        bases, then 3'_FLANK is also required.
                                        Field may be up to 255 b.p. in size.
                                        If greater than 255 b.p., excess
                                        characters should be put in 3'_FLANK.
3'_FLANK:   <sequence>                  Flanking sequence 3' of the assayed
                                        region. Field is required if the
                                        3'_ASSAY is less than 25 b.p. or if
                                        the 5'_ASSAY and 3'_ASSAY fields combined
                                        are less than 100 b.p. Minimum b.p. requirement
                                        is suspended if a valid SNP_LINK field
                                        is provided. White space allowed, and
                                        will be ignored.
||
```

EXAMPLES

For a Submission for the Whitehead Institute, given the handle, 'WI', a submission of a set of SNP assay might look like:

```
TYPE:SNPASSAY
HANDLE:WI
BATCH: 1.98
MOLTYPE:Genomic
METHOD:RESEQ
SYN NAMES:WI-SNP,DnaId,MapDna
COMMENT:
Here is where some public comment that applies to the entire
batch of SNPS could be put.
PRIVATE:
Here is where a note to NCBI regarding processing that would
not be seen by the outside, could be put.
Note that these are is not exactly real SNPs, as
the data were modified.
||
SNP:WI|WIAF-1234567
SYNONYM:EST4291092,EST8291092,EST7291092
ACCESSION:H30533
LENGTH:101
5'_ASSAY:GGCAGGGAAGGAAAATCCTAGGGNCAGCATTGGGGAGGGGGGGACTCTG
OBSERVED:C/T
3'_ASSAY:TAAATTTATTGGGCAACAGGCTGCAGGTGAGGGGGCTGACAGGAGGAGGGA
||
SNP:WI|WIAF-1722
SYNONYM:STS-T17494,STS-T17494,STS-T17494
ACCESSION:T17494
LENGTH:269
5'_FLANK:CTTTCCCTCATCCCCTCTTCCACCACACCATCCCGGAACAAGTGCTCCAGGATT
5'_ASSAY:CCCTGCCCACTGGCCATTTTGGAGTGTGTCC
OBSERVED:A/T
3'_ASSAY:GTGGGTAGCAATGTGGAAACCACCAGGGCCTTTGTGGAGAAAA
3'_FLANK:TGGAGGGGGTTGAGGGAGTCCCAGGAGGGGCTTATTTGAGGGCCTTTGCCACTT
   GCTCATAGGCGAGCTCGATCTCCTCATCATCTGGACAGGTGGAAGCGAATTCTT
   CCCGGGCGTAGGCATTGCTCAAGTACCGAT
||
```

Back to Table of Contents

---

dbSTS submission elements

See the web page for details on submissions to dbSTS, which may come through the dbSNP channel for simultaneous submissions. There are seven types of deliverable sections which will be passed on to dbSTS for simultaneous submission to dbSTS and dbSNP:

a. Publication
b. Source
c. Contact
d. Protocol
e. Buffer
f. STS
g. Map

Of these, only the Publication and Contact types are shared in submissions to dbSTS and dbSNP. (So these two are the only ones also detailed in this page.) Note the addition of the *handle* to the Contact section.

Note:

Data sections that are for STS but not SNP, such as buffers and protocols should not be submitted UNLESS there is an STS submission being done simultaneously. If data are not available for some fields, the field can either be omitted entirely, or the tag may be included with an empty data field. Please do not put '*', "-", etc to indicate missing data. Handle and local id spelling must be completely identical for matching. Similarly, the citation information must match the title of a Publication section, exactly. dbSTS uses the full *Contact name* for matching, while dbSNP uses the shorter handle. So for simultaneous STS/SNP submission, care must be taken with both. If you wish, you can submit sources, pubs, contacts, protocols, buffers, methods, populations, STS, and SNPs all in one file - the TYPE field will differentiate them for the parsing software. However, if you are submitting new sources, protocols, buffers, contacts,methods, populations and/or publications in the file with SNPs, and the new SNPs refer to them, they must precede the SNPs in the file, otherwise the SNP crossmatching will not succeed.

Back to Table of Contents

SNP Submission of Genotype data

SUBMITTER BEWARE!
Care must be taken when describing allele frequencies and genotypes

Genotype submissions require the specification of an allele's strand with respect to the **<snp_id>** field of the submission. We have also updated the submission format to accept reference clustered (rs) SNPIDs.

The **specification of the strand field is necessary to** ensure proper calculation of allele frequencies across multiple submissions.

The **following example** outlines the basic problem.

Consider the case of a SNP (an A/T polymorphism colored red) shown below in double stranded sequence:

5'-GATTAGTAA/TGCCGAGCTG-3' --> Forward strand
3'-CTAATCATT/ACGGCTCGAC-5' <-- Reverse strand

One submitter reports the frequency of the alleles observed with regard to the forward strand as:
A = .25 and T= .75

A second submitter reports the frequency of the alleles observed on the reverse strand as:
A = .75 and T=.25

Without strand information, **these results would appear to contradict each other because** an observer would make **the assumption** that both submitters were **reporting** allele frequencies **with respect to the forward strand. When strand is taken into consideration**, it is apparent to an observer that these two submitters are reporting equivalent allele frequencies.

Because the potential discrepancy detailed in the above example, dbSNP NOW REQUIRES SUBMITTERS TO SPECIFY ORIENTATION USING A NEW FIELD CALLED <STRAND>.

The values for <STRAND> are:

Value	Description
[SS_STRAND_FWD]	The alleles for this submission are the nucleotides that occur on the same strand as the 5' and 3' flank of the ss specified in the <snp_id> field.
[SS_STRAND_REV]	The alleles for this submission are the reverse complement of the nucleotides that occur on the same strand as the 5' and 3' flank of the ss specified in the <snp_id> field.
[RS_STRAND_FWD]	The alleles for this submission are the nucleotides that occur on the same strand as the 5' and 3' flank of the rs specified in the <snp_id> field.
[RS_STRAND_REV]	The alleles for this submission are the reverse complement of the nucleotides that occur on the same strand as the 5' and 3' flank of the rs specified in the <snp_id> field.

Below are three example submissions showing how the new <STRAND> field is used.
Orientation will be specified using the <STRAND> field in the SNPPOPUSE (for allele and/or genotype frequencies) and SNPINDUSE (for individual genotypes) submission sections.

For a more detailed explanation of how to determine which value to use in the <STRAND> field is available.

Genotype Submission Examples:
FREQUENCY DATA: The frequency and count examples below illustrate how to report sample estimates of allele frequency, genotype frequency, and measures of observed heterozygosity.

SNP Use on Individuals Sections

At The Beginning of the section describing the SNP assays there is a header that supplies information that applies to the rest of the section. The required fields in this header are HANDLE, BATCH, and METHOD. Two formats are provided for the repeating data within this section. In the first case genotype data is grouped on indivudal ID, and in the second case the data are grouped on SNP ID. The second format is useful when multiple SNPs have unique METHOD_EX lines in the header. One format must be used consistently within a single batch. Separate batches may use different formats.

```
TYPE:                                   Entry type, must be "SNPINDUSE" for these.
HANDLE:   <handle>
BATCH:       <local batch id>           The submitter will name each batch for
                                        ease of communication.  This name will
                                        only be unique for a particular
                                        submitter.
METHOD:      <local method Identifier>
METHOD_EX:   Free text                  variation from given method
CITATION:    Title of publication       To match the title of an entry in a
                                        publication section of this submitter.
COMMENT:     Free text                  for public, will be shown with each SNP
                                        assay in this batch.
PRIVATE:     Free text                  for NCBI to aid in processing
||
     - - - - - Repeating for each Individual [FORMAT 1] - - - - -
ID:           <handle>|<local population identifier>:<local individual Identifier>
SNP:          <SNP ID> :observed allele[/allele]|<strand>
                                        Of course only two alleles make sense
                                        in this context, unless individual is
                                        triploid for the SNP locus.   So a second
                                        variation may be repeated after a slash.
[SNP:         more SNPs, if multiple SNPS assayed in this individual]
```

GENOTYPE DATA: (Uses the NIH Polymorphism Discovery Resource "NIHPDR" as population sample.)

```
TYPE:SNPINDUSE
HANDLE:WHOEVER
BATCH:1-98
METHOD:MYMETHOD
||
ID: NCBI|NIHPDR:1
SNP:NCBI|rs1:A/T|RS_STRAND_FWD <---- SNP identified by dbSNP accession
SNP:WI|WIAF-1722:G/C|SS_STRAND_FWD <---- SNP identified by submitter's local ID
SNP:NCBI|ss13:-/G|SS_STRAND_FWD <---- heterozygous individual with genotype
SNP:NCBI|rs101:C/C|RS_STRAND_FWD <---- homozygous individual with genotype
SNP:WI|999:115:(homozygous)|SS_STRAND_FWD <---- homozygous individual without genotype
SNP:WI|1001:(indeterminate)|SS_STRAND_FWD <---- no data
||
ID:NCBI|NIHPDR:2
SNP:NCBI|rs1:A/A|RS_STRAND_FWD
SNP:WI|12345:G/C|SS_STRAND_FWD
SNP:NCBI|ss13:A/G|SS_STRAND_FWD
SNP:NCBI|rs101:G/C|RS_STRAND_FWD
SNP:WI|999:115:T/T|SS_STRAND_FWD
SNP:WI|1001:(indeterminate) |SS_STRAND_FWD
||
   - - - - - Repeating for each Individual [FORMAT 2] - - - - -
SNP:        <SNP ID>
ID:         <handle>|<local population identifier>:<local individual Identifier> :observed allele[/allele];
                          Like above, only two alleles make sense
                          in this context, unless individual is
                          triploid for the SNP locus.   So a second
                          variation may be repeated after a slash.
[ID:        more individuals, if multiple people have been assayed for this SNP]
||
EXAMPLE (Assumes a global population, "NIH_PANELA" and two SNPs typed with different restriction enzymes)
TYPE:SNPINDUSE
HANDLE:WHOEVER
BATCH:1-2002
METHOD:RESTRICTION_ENZYME
METHOD_EX: ECO_RI
||
SNP:NCBI|rs1|RS_STRAND_FWD
ID:NCBI|NIHPDR:1:A/T
ID:NCBI|NIHPDR:2:-/G
ID:NCBI|NIHPDR:3:A/A
ID:MYPOP1:1:C/C
ID:MYPOP1:2:C/T
ID:MYPOP2:1:A/-
||
SNP:WI|WIAF-1722|SS_STRAND_FWD
ID:NCBI|NIHPDR:1:G/C
ID:NCBI|NIHPDR:2:G/G
ID:NCBI|NIHPDR:3:G/G
ID:MYPOP1:1:A/A
ID:MYPOP1:2:A/A
ID:MYPOP2:1:T/T
||
```

Back to Table of Contents

SNP Use on Populations Sections

At the beginning of the section describing the SNP assays there is a header that supplies information that applies to the rest of the section. The required fields in this header are HANDLE, BATCH, and METHOD.

Population variation information can now be submitted in three classes: ALLELE frequencies, GENOTYPE frequencies, or OBSERVED HETEROZYGOSITY. Multiple classes of data may be submitted for the same population. The keywords SNPFREQ: and SNPCOUNT: have been replaced by ALLELEFREQ: and ALLELECOUNT: as noted in the guidlines below.

```
TYPE:                                   Entry type, must be "SNPPOPUSE" for these.
HANDLE:     <handle>
BATCH:              <local batch id>          The submitter will name each batch for
                                        ease of communication.  This name will
                                        only be unique for a particular
                                        submitter.
METHOD:      <local method Identifier>
METHOD_EX:   Free text                  variation from given method
CITATION:    Title of publication       To match the title of an entry in a

                                        publication section of this submitter.
COMMENT:     Free text                  for public, will be shown with each SNP
                                        assay in this batch.
PRIVATE:     Free text                  for NCBI to aid in processing
||
     - - - - - Repeating for each Population - - - - -
ID:           <handle>|<local population identifier>
SAMPLESIZE:   <number>                  How many in sample (population) REQUIRED
                                        The units should be number of chromosomes.
- - - To report ALLELE FREQUENCIES use ALLELEFREQ or ALLELECOUNT - - -
ALLELEFREQ:  <SNP ID>:<allele>=<frequency>[/<allele>=<frequency>/...]|<strand> to report a frequency for each allele
ALLELEFREQ:  <SNP ID>:<allele>=<lo_frequency>-<hi_frequency>[/<allele>=<lo_frequency>-<hi_frequency>/...]|<strand>
                                        to report a frequency range (lo_frequency,hi_frequency)
                                        for each allele

ALLELECOUNT: <SNP ID>:<allele>=<count>[/<allele>=<count>/...]
                                        to report allele frequency as an integer fraction of SAMPLESIZE
                                        See variation, above, for how to report
                                        a variation.  Of course multiple alleles
                                        make sense in this context.
- - - To report GENOTYPE FREQUENCIES use GENOTYPEFREQ or GENOTYPECOUNT - - -
GENOTYPEFREQ:  <SNP ID>:<genotype>=<frequency>[/<genotype>=<frequency>/...]|<strand>
                                        to report a single frequency for each genotype
GENOTYPEFREQ:  <SNP ID>:<genotype>=<lo_frequency>-<hi_frequency>[/<genotype>=<lo_frequency>-<hi_frequency>/...]|<strand>
                                        to report a frequency range (lo_frequency,hi_frequency)
                                        for each genotype

GENOTYPECOUNT: <SNP ID>:<genotype>=<count>[/<genotype>=<count>/...]|<strand>
                                        to report genotype frequency as an integer fraction of SAMPLESIZE
                                        multiple genotypes make sense in this context.
- - - To report OBSERVED HETEROZYGOSITY use HETFREQ or HETCOUNT - - -
HETFREQ:  <SNP ID>:(heterozygous)=<frequency>/(homozygous)=<frequency>
                                        to report a single frequency for each genotype
HETCOUNT: <SNP ID>:(heterozygous)=<count>/(homozygous)=<count>
                                        to report heterozygosity as an integer fraction of SAMPLESIZE
TYPE:SNPPOPUSE
HANDLE:WHOEVER
BATCH:1-2002
METHOD:MY_FREQUENCY_METHOD
||
ID:MYPOP|MYSAMPLE1
SAMPLESIZE:100
ALLELEFREQ:NCBI|ss1:A=0.50/T=0.50|SS_STRAND_FWD   << reports allele frequency using "ss" notation
ALLELECOUNT:WI|12345:G=30/C=70|SS_STRAND_FWD   << reports frequency using submitter notation
ALLELECOUNT:NCBI|ss3:C=100|SS_STRAND_FWD     <<<< reports no variation in this sample
ALLELECOUNT:WI|1001:(indeterminate)=50/A=25/T=25|SS_STRAND_FWD     <<<< reports missing data
ALLELEFREQ:NCBI|ss10:T=0.05-0.15/C=0.85-0.95|SS_STRAND_FWD  <<<< reports a frequency range for each allele
GENOTYPEFREQ:NCBI|ss5533:AA=0.5/AC=0.3/CC=0.2|SS_STRAND_FWD  <<<< reports frequency for each genotype
HETCOUNT:NCBI|ss6201:(heterozygous)=5/(homozygous)=45  <<<< reports heterozygosity for locus strand not required
||
```

DETERMINING THE VALUE TO ENTER IN THE <STRAND> FIELD : *For this example, consult table 1 (below) for descriptions of the SNP_IDs used.*

Consider a dsDNA sequence representing the submitted SNP **ss3348464**, an A/G variation as shown below with the polymorphism in red:

ss3348464: CTTTCGTTAGGCTAGTTA/GGCTGAGCCATTGTATG

However, **ss3348464** clusters with other SNPs to make **rs3325**, which is the same variation as **ss3348464**, only defined on the other strand as a C/T variation as shown below with the polymorphism in red.

rs3325: CATACAATGGCTCAGCT/CAACTAGCCTAACGAAAG

Now consider a lab that uses sequence specific oligonucleotide (SSO) hybridization to detect the SNP at **ss3348464** (or, in reverse complement, **rs3325**). This lab may choose to design SSO probes from the forward ss strand, the reverse ss strand, the forward rs strand, or the reverse rs strand. Typically, two probes designed from one strand are assayed against a sample and evaluated as positive or negative for hybridization.

The <strand> field is used in the submission to define the strand (and hence precise allele) from which the SNP allele-assay is evaluated. <Table 1 (below) illustrates possible probe sequences that can be developed on either strand, the SNP alleles they genotype, and the strand field value that should be used in the submission.

Table 1: strand designations for possible probe configurations for **ss3348464 or rs3325**

SNP_ID USED IN SUBMISSION	PROBE SEQUENCE	ALLELE REPORTED WHEN POSITIVE	STRAND FIELD VALUE
ss3348464	TGGCTCAGCTAACTAGCCT	A	SS_STRAND_FWD
ss3348464	TGGCTCAGCCAACTAGCCT	G	SS_STRAND_FWD
ss3348464	AGGCTAGTTGGCTGAGCCA	C*	SS_STRAND_REV
ss3348464	AGGCTAGTTAGCTGAGCCA	T*	SS_STRAND_REV
rs3325	AGGCTAGTTGGCTGAGCCA	C*	RS_STRAND_FWD
rs3325	AGGCTAGTTAGCTGAGCCA	T*	RS_STRAND_FWD
rs3325	TGGCTCAGCTAACTAGCCT	A	RS_STRAND_REV
rs3325	TGGCTCAGCCAACTAGCCT	G	RS_STRAND_REV

* Nucleotide state tested by these probe sequences are reverse complement to the alleles defined for the specific submission ss3348464 and are the same as the alleles defined for the rs3325.

NOTE: Defining <strand> is particularly important in cases where the observed SNP alleles are complimentary such as in a G/C or A/T polymorphism.

Back to Table of Contents

SNP Assay updates (to be used to update validation data on existing submissions)

TYPE: BATCH_UPDATE	To set a success rate for a batch with a previously unclassified success rate. This can potentially involve adding a population
HANDLE: <handle>	
BATCH: <local batch id>	ID if it was unreported when the batch was initially submitted.
NEW_METHOD: <local method id>	To change the method used for the bath. Use VALIDATION section for adding additional methods to SNPs. Use of NEW_METHOD will require database administrator confirmation at load time to ensure data integrity.
SUCCESS_RATE: <percentage>	
POPULATION: <population id>	
COMMENT:	
LINKOUT_URL:	
||	
TYPE: BATCH_REASSIGN	To change the batch id with which a SNP is associated. Used to move SNPs to a newly created batch ID with a different SUCCESS_RATE and/or population. Method ID and method_count for SNPs remain the same. New batch inherits all properties from Old batch except for NEW_SUCCESS_RATE, POPULATION, COMMENT and LINKOUT_URL as noted.
HANDLE: <handle>	
OLD_BATCH: <local batch id>	
NEW_BATCH: <local batch id>	
NEW_SUCCESS_RATE:<percentage>	
NEW_POPULATION:<population id>	
COMMENT:	
LINKOUT_URL:	
||	

- - - - - Repeating for each SNP to be reassigned- - - - -

SNP: <ID, ss# or local>	ID must exist in OLD_BATCH
||	
TYPE: VALIDATION	Used to add a new method to a SNPs method history. The new batch will inherit

	properties from Old Batch except for NEW_SUCCESS_RATE, POPULATION, and NEW_METHOD as noted.
HANDLE: <handle>	
OLD_BATCH: <local batch id>	
NEW_BATCH: <local batch id>	
NEW_SUCCESS_RATE:<percentage>	Probability that SNP is real, based on validation data.
NEW_METHOD: <local identifier>	ID of validation method used on a set of optional variation from given method
SNPs METHOD_EX:	Free text
NEW_POPULATION:<local identifier>	
COMMENT:	optional comment
LINKOUT_URL:	optional linkout to submitter website
||	
- - - - -Repeating for each SNP that was validated - - - - -	
SNP: <ID> (SEGREGATES=YES|NO|UNKNOWN; HOMOZYGOTE_FOUND=YES|NO|UNKNOWN)	Update of SEGREGATES and HOMOZYGOTE_FOUND is optional
||	

SNP Withdrawn

TYPE: WITHDRAWN	Used to mark a SNP as withdrawn. SNP will retain ss#, but type will change from SNP to WITHDRAWN (WD).
HANDLE: <handle>	
||	
EVIDENCE: GeneDuplication	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	
EVIDENCE: Artifact	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	
EVIDENCE: NotSpecified	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	
EVIDENCE: AmbiguousMapLocation	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	
EVIDENCE: LowMapQuality	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	
EVIDENCE: DuplicateSubmission	
- - - - - Repeating for each SNP withdrawn for this reason - - - - -	
SNP: <ID, ss# or local>	
||	

dbSNP Reports

Submitted SNP reports

The report for each SNP will consist of the header and SNP-specific data from the original submission file. The following data validation fields will be computed by NCBI if the necessary data are present, and may also be present in the ASSAY section:

```
HW_PROB:        <computed by NCBI>          Chi-square probability computed from
                                        use on individuals section if present.
HET:            <computed by NCBI>      Estimated heterozygosity for the locus,
                                        computed from use on individuals section,
                                        if present.
QA_STATUS:      <computed by NCBI>      Summary index of above validation-related
                                        fields. Integer valued.
```

Back to Table of Contents

Changes in version 2.0 (September 2002)

- SPEC (STRAND): STRAND Field added to capture orientation information on genotype and allele frequency submissions.

Changes in version 1.05 (September 9, 1999)

- SPEC (NOVARIATION): Keyword "COMMENT" added to the repeating STS/ACCESSION sections for free text annotation.
- SPEC (SNPASSAY): Keyword "POPULATION" added to batch header to report population used for marker discovery.

Changes in version 1.04 (August 5, 1999)

- SPEC (NOVARIATION): No Variation section defined for reporting survey results with no polymorphism detected.
- SPEC (SNPASSAY): Keyword "ANCESTRAL" added to report ancestral allele when known.
- SPEC (SNPASSAY): Keywords "STRAIN" and "CULTIVAR" added to support submissions on laboratory strains of plants and animals.
- SPEC (SNPASSAY): "LINKOUT_URL" keyword added to link reports back to submitter website.
- SPEC (SNPASSAY): "SAMPLESIZE" keyword is now required. It can be specified in the batch header to give a default sample size for the entire batch, and used in the repeating "SNP" section to give an override value for particular records.
- SPEC (SNPASSAY, SNPINDUSE, SNPPOPUSE): The keyword "SUBMITTER" changed to "HANDLE" for consistency with other sections.
- SPEC (SNPPOPUSE): Population frequency data may be submitted for *allele frequencies*, *genotype frequencies*, and *observed heterozygosity*. Range estimates added as an optional format for frequency data.
- EX (SNPINDUSE): Revised example of submitting genotype data from a common resource (NIH Polymorphism Discovery Resource, NIHPDR).

Changes in version 0.14 (November, 1998)

- A new section describing the organization of data in dbSNP and the accession numbers for dbSNP objects and records.
- PROCEDURE: Database administration and policy statements have been grouped together. A note has been added to alert submitters of the differences in Hold Until Published (HUP) Policies between dbSTS and dbSNP.
- PROCEDURE: Updating entries in dbSNP is now defined.
- PROCEDURE: Single literature citations in a batch will be associated with assay reports in same batch if a citation field is absent from an assay batch header.
- SPECIFICATION: Two kinds of flanking sequence fields are supported: 5' and 3' *assay sequence* which should be used to indicate the sequence observed in a specific experimental method, and 5' and 3' *flanking sequence* which should be used to denote known sequence that surrounds the polymorphism, but is undetected by the experimental method.
- SPECIFICATION: Flanking sequences may include IUPAC ambiguity characters to denote sites with known variation.
- SPECIFICATION: Several optional validation and information fields have been added to the assay section.
- SPECIFICATION: A format for specifying microsatellite variation has been introduced.
- SPECIFICATION: Two formats are now supported for the repeating section in USE ON INDIVIDUALS: the original format and an alternative format with SNP and ID fields reversed. The second format was introduced to accommodate batch file processing of cases where each SNP is associated with a unique METHOD_EX line, as in the case of restriction enzyme-based assays.

Back to Table of Contents

Database components currently under construction

1. Establishing a ftp version of the entire database.
2. Implementing a web-based SNP submission interface for small batches of submission.
3. Definition of an NCBI feature entry for dbSNP data. This will be used to annotate SNP data on other NCBI resources.
4. A format for sample ascertainment conditions in reports of snp discovery.
5. An extension to the database to accommodate haplotype data.

Back to Table of Contents

Locus validation fields

Data validation can be maintained for both submitted assay reports and abstracted SNP objects. At the level of an individual submitted assay report, dbSNP provides the following fields to assess the quality of the data:

- **Report linking** to other submitted SNP assays. The SNP_LINK field can be used to associate a submission with another record in the databse. This field can be used by itself to link a new assay for a particular SNP with other reports for the SNP in the database, or it can be used in conjunction with the METHOD_FAILURE comment field (below) to report problems with existing assays or primer pairs.
- **Method_Failure:** explicit text field to report problems with a SNP. Used in conjunction with the SNP_LINK field.
- **Mendelize:** whether the SNP conforms to Mendelian inheritance
- **Individual homozygosity:** if homozygotes were observed. Excessive heterozygosity may be the product of paralogous sequence organization, rather than
- **Hardy-Weinberg chi-square value:** calculated from use on individuals section. It is commonly used to test for the departure of genotype frequencies from their expected values in a neutrally evolving population.
- **Heterozygosity:** a common measure of genetic diversity at the locus.
- **PCR confirmed:** to indicate that polymorphism was detected in multiple products and not simply a potential artifact
- **frequency data:** will report variation by population or individual as submitted. These data can be submitted at a later time by any lab who performs experiments with the SNP.
- **validation status:** numeric field computed as a quality score to summarize validation data if present.

Abstract SNP records have validation fields that summarize the QA data for each of the submitted SNP reports they encompass.

- **Mendelize**: if the SNP conforms to Mendelian inheritance in any report.
- **Individual homozygosity**: if homozygotes were detected in any report.
- **Hardy-Weinberg chi-square value**: best, worst, and average value. Calculated from use on individuals sections when provided.
- **Individual consistency**: error if individual genotypes vary across reports. Computed from use on individuals sections.
- **Validation status**: fuzzy logic field to summarize above information into a single numerical value.

Comments or Questions?
Write to the NCBI Service Desk